Speech spectral envelope estimation throught explicit control of peak evolution in time

Elizabeth Godoy, Olivier Rosec, Thierry Chonavel

SPEECH SPECTRAL ENVELOPE ESTIMATION THROUGH EXPLICIT CONTROL OF PEAK EVOLUTION IN TIME

Elizabeth Godoy [1]*, Olivier Rosec*[1]*, Thierry Chonavel* [2]

[1] Orange Labs, Lannion, France
[2] Telecom Bretagne, Signal & Communication Department, Brest, France
{elizabeth.godoy,olivier.rosec}@orange-ftgroup.com,thierry.chonavel@telecom-bretagne.eu

ABSTRACT

This work proposes a new approach to estimating the speech spectral envelope that is adapted for applications requiring time-varying spectral modifications, such as Voice Conversion. In particular, we represent the spectral envelope as a sum of peaks that evolve smoothly in time, within a phoneme. Our representation provides a flexible model for the spectral envelope that pertains relevantly to human speech production and perception. We highlight important properties of the proposed spectral envelope estimation, as applied to natural speech, and compare results with those from a more traditional frame-by-frame cepstrum-based analysis. Subjective evaluations and comparisons of synthesized speech quality, as well as implications of this work in future research are also discussed.

Keywords: speech analysis, spectral envelope, voice conversion

1. INTRODUCTION

Currently, the spectral envelope of speech signals is most commonly estimated frame-by-frame, using Linear Prediction (LP) or cepstrum-based methods to generate parameters such as Line Spectral Frequencies (LSF) or cepstral coefficients, respectively [1]. These techniques yield high quality results for analysis-synthesis applications, speech coding, and for contexts using prosodic modifications of speech [2]. Unfortunately, this approach to treat frames independently can cause problems and yield poor synthesis quality when using time-varying modifications of the spectral envelope [3]. One important application requiring this type of modification is that of Voice Conversion (VC), which aims to modify the utterance of an individual (source) speaker so that it sounds as if a different (target) speaker uttered the same phrase. This process involves three main stages: first, modeling and analysis of acoustic parameters to be converted (e.g. spectral envelope); second, training through learning a mapping between the source and target parameters; third, transformation of the source parameters. In essence, since the estimation of the spectral envelope in the analysis stage treats speech frame-by-frame there can be a loss of coherence in the signal when transformation tries to take into account spectral variations in time, across a sequence of frames. Current approaches to VC try to maintain coherence in the converted signal by adapting modification algorithms to incorporate time-evolution of spectral parameters in speech, as in [4], for example. In seeking to adapt only methods of modification, these works concentrate uniquely on the training and transformation stages of a VC system. We argue that it is necessary to consider the temporal evolution of the spectral envelope directly in the acoustic analysis stage of conversion. Accordingly, in this work, we adapt estimation of the spectral envelope to incorporate interdependence of frames in localized speech segments, namely within phones, in order to model evolution of the spectral parameters in time. The underlying idea is then to optimize estimation of the spectral envelope over a sequence of frames rather than within an individual frame.

In our adapted acoustic analysis, we represent the spectral envelope as a sum of Gaussian peaks and we model the evolution of these peaks within a phoneme. This choice of representation for the spectral envelope has several interesting attributes in the VC context. First, this representation is perceptually relevant because peaks in the spectrum correspond generally to formants, which are key features in human speech perception [3]. Second, the peak parameters, namely location, amplitude and bandwidth, are localized in frequency and can be treated independently, two properties which enable flexibility and a higher degree of control in modification of the spectral envelope, as discussed in [4]. Third, our choice of spectral representation enables more natural modeling of the temporal evolution of spectral parameters. Specifically, peaks in the spectrum are related to formants or, more generally, resonances of the vocal tract. In speech production, for the majority of sounds, particularly vowels (which contain the most speaker-dependent information), the vocal tract shape does not change abruptly within a phone. Therefore, peaks in the spectrum generally evolve smoothly in time. With our localized parameterization of peaks that limits interference between

different regions of the spectrum, we can separately track and control peak parameters in time and ensure smooth evolution of spectral parameters within a phone. In sum, we propose a model for acoustic analysis of the spectral envelope that is particularly well adapted for VC in that it provides parameters linked to speech perception that are localized in frequency and can be followed and controlled in time.

In the following sections, we will discuss our novel approach to estimating the spectral envelope, based on modeling the temporal evolution of spectral peaks across a phoneme. We will explain implementation of the acoustic analysis and we will then show examples of our model applied to natural speech. Additionally, we will compare properties of our proposed modeling with those of a more traditional cepstrum-based frame-by-frame approach to modeling the spectral envelope. Finally, we will discuss qualitative evaluations of our results followed by conclusions and implications of this work in future research.

2. SPECTRAL ENVELOPE REPRESENTATION WITH TIME-EVOLVING PEAKS

2.1. Modeling Spectral Peaks in a Frame

We will first describe modeling peaks in the spectral envelope within an individual frame. We model the spectral envelope for frame i as a sum of M_i Gaussians:

$$S_i(f) = \sum_{m=1}^{M_i} a_m^i N(f; \mu_m^i, \sigma_m^i)$$

where $S_i(f)$ is the spectral envelope magnitude and f represents frequency. The parameters $\{a_m^i, \mu_m^i, \sigma_m^i\}$ respectively represent the amplitude, location and variance of the Gaussian-modeled peak m in frame i. Unlike the work in [5] and [6], we model spectral peaks using a non-parametric representation for the speech spectrum, namely the Discrete Fourier Transform (DFT). In this way, we avoid alterations to spectral characteristics that can be introduced by parametric modeling, e.g. with LP or cepstral analysis. The amplitude and location of the spectral peaks are taken from pick-peaking on the DFT using a frequency mask with size depending on both location in the spectrum and fundamental frequency of the frame. These dependencies respectively allow us to both achieve higher resolution in low frequency parts of the spectrum (in which human hearing is more sensitive) and to avoid modeling harmonic peaks. The number of Gaussian peaks is limited by the parameters of the frequency mask, but is not fixed for each frame. However, we allow a maximum of 20 peaks per frame. Once the peak locations and amplitudes are determined, we locally calculate the variance of each peak considering the spectrum around only neighboring peaks. In this way, we avoid interference between different parts of the spectrum in our parameter calculation. Specifically, in-between two peaks, we find a point in the spectrum that is equiprobably generated from both surrounding peaks. We then calculate the sample variance for the left and right peaks using the regions to the left and right of this point, respectively. The variance for a single peak is then the average of the sample variances calculated on both sides. Finally, since human hearing is highly sensitive to variations in the low-frequency part of the spectrum, we limit variability of the spectral envelope around the first peak. In particular, we keep the spectral amplitude constant up to the first peak maximum and we smoothly interpolate between the first and second peak amplitudes.

Figure 1 shows a plot of the DFT of an individual frame with the spectral envelope obtained from our Gaussian-peak parameters. We have also included the spectral envelope generated from the discrete cepstrum coefficients as calculated in [7], with an order of 40. As can be seen in Figure 1, both approaches capture detailed variations in the spectral envelope shape. In the dashed-curve generated from the cepstral analysis, there is high accuracy in the envelope modeling for low frequencies since a bark scale is used in calculating the coefficients. In our analysis, the resolution of envelope modeling can be adapted to different regions of the spectrum by changing constraints on the frequency mask accordingly. The most significant differences between the two approaches is that, while the shape of the envelope generated from cepstral coefficients is not constrained and can thus yield significant variations (e.g. rolling curve, rapid fluctuations) within a frame or between frames, our proposed model represents the envelope explicitly as a combination of peaks that relate more closely to human speech perception. Thus, we sacrifice some accuracy in spectral estimation within a particular frame in order to better model parameter evolution across a sequence of frames.

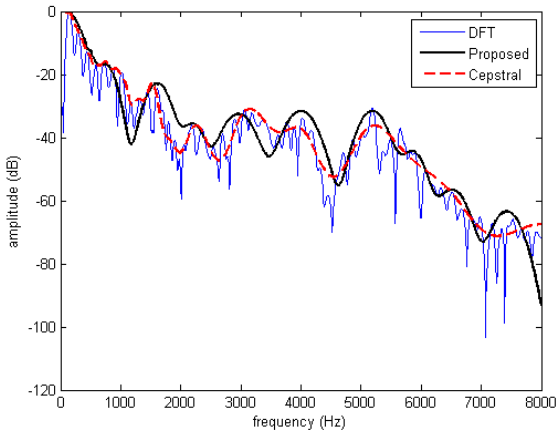

Figure 1. DFT (blue) of a speech frame with normalized spectral envelopes generated from Gaussian peak parameters (solid, black) and from discrete cepstral coefficients (dashed, red).

2.2. Time-Evolution of peaks in a Phoneme

One key aspect of our approach comes from incorporation of phonetic segmentation and temporal context directly into the analysis of the spectrum of a frame. Specifically, we take the approach described above for Gaussian-peak modeling of an individual spectral envelope and we apply constraints such that the peak evolutions over time, within a phoneme, are smooth. The underlying assumption in this analysis is that the peaks do not move drastically between two adjacent frames within the same phone, which is the case for the majority of voiced phones. The following section describes our acoustic analysis in more detail.

We first consider the "stable" part of a voiced phoneme, comprised of the three center frames, where the frames are determined pitch synchronously. The center of a phoneme is the region in which there is the least variation between frames and the smallest influence of adjacent phonemes. Thus, this region is generally the most reliable in representing frames of the phoneme. Accordingly, our approach is to anchor our analysis around this region. Specifically, we begin by analysing the three center frames of the phoneme individually. We select the two fames with spectral envelopes that are the closest and we align the peaks between these frames. This alignment is based on locally minimizing the differences between peak locations of the two frames. We average the parameters of the aligned peaks to generate the parameters of the center frame of the phoneme. From this stable frame, we then analyse frames sequentially, moving outward from the center (to the left and right), towards the phoneme boundaries. We constrain pick-peaking on the outward frame so that the peaks fall within a local region of peaks of the adjacent inner frame. We then align the peaks between the inner and outer frames, using the same alignment described above, and set the outer frame parameters to the average of the aligned-peak parameters. This method of analysis thus ensures smooth evolution of the peak parameters in time, within a phoneme, while ultimately placing more emphasis on the stable part of the phone.

3. ANALYSIS RESULTS FOR NATURAL SPEECH

We evaluated our method for acoustic modeling of the spectral envelope on natural speech taken from corpora used in France Télécom's Text-to-Speech synthesis system Baratinoo. The speech is sampled at 16kHz and is segmented into phones, labelled, and pitch-marked. In the following section, we will highlight some important properties of our method of acoustic analysis and contrast the results with those from the frame-by-frame discrete cepstral analysis of the same speech.

Figure 2 shows the spectrograms generated from the Gaussian-peak parameters and the discrete cepstral coefficients, in the upper and lower plots, respectively. Superimposed on these spectrograms are circles showing the peak locations calculated using the techniques described in section 2. Peak locations in the stable center frames are filled in. First, we see in this example that we are able to successfully track the evolution of high-energy regions of the spectrum represented in both analyses. Unlike traditional approaches to tracking spectral peaks, this tracking is incorporated directly into the spectral analysis. Second, we note that in the spectrogram from our proposed analysis, the spectral evolution within a phone is smooth and distinct regions of the spectrum are clearly distinguished in time. The correlation between adjacent frames in a phone is clear across all frequencies. On the other hand, in the frame-by-frame approach, we see notable variations across the spectrum frequencies between adjacent frames, even within a vowel, as seen in the 'I'.

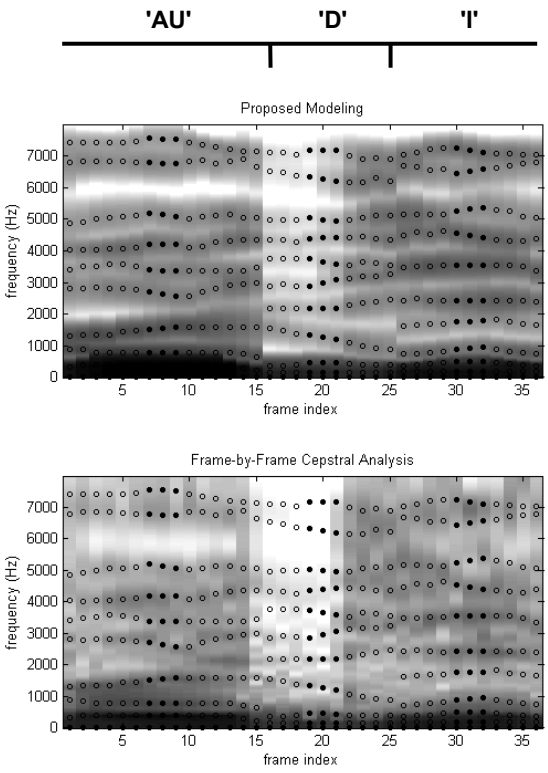

Figure 2. Spectrograms of the sequential phonemes 'AU'-'D'-'I' from a French phrase. Peak locations calculated from our proposed analysis are indicated in both plots by black circles (filled indicate stable center frames).

Figure 3 focuses our analyses within an individual phoneme. We show 3-dimensional plots of the amplitude-normalized spectral envelopes for all frames in a particular phoneme, concentrating on the lower half of the spectrum. In these plots, we see that our acoustic analysis modeling the peak evolution over time is able to capture spectral trends within the phoneme in a regular and coherent way. On the other hand, frame-by-frame analysis of the phoneme yields sporadic and discontinuous variations across the spectral envelopes. These properties will become significant in training and transforming spectral features in a conversion context. In particular, in the training stage, spectral envelopes will be grouped

according to their similarity. In the case where adjacent envelopes have distinctly different features, transformation could then apply a significantly different function to each frame, which may create incoherence in the converted signal. By contrast, our approach seeks to avoid this problem by enforcing regularity within a phoneme.

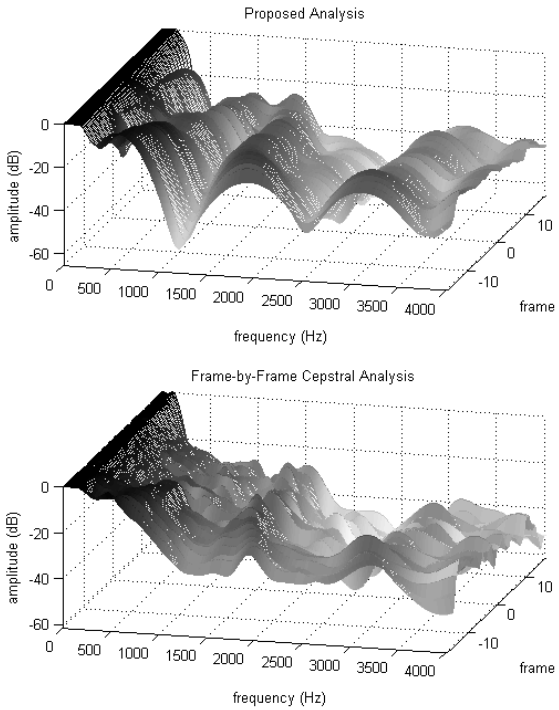

Figure 3. 3-D plots of the amplitude-normalized spectral envelopes from an instance of the phoneme 'A.' Frames are shown in sequence wrt the phoneme center at 0.

Finally, in this initial evaluation of our proposed method for spectral analysis, we performed analysis-synthesis on selected phrases from different speakers in the France Télécom corpora (two male and two female). Specifically, we used a Harmonic plus Noise Model (HNM) for the speech, as described in [7], and we obtained the harmonic amplitudes from sampling the spectral envelopes generated by our proposed analysis and by the discrete cepstral analysis. The original harmonic phases are kept in both cases. In informal listening tests, we noted a slight degradation in the synthesis quality using our proposed approach, while the cepstral analysis yielded a quality close to that of the original phrase. This degradation can be expected, as we sacrifice some accuracy in representing the spectral envelope of an individual frame in order to better represent the evolution of spectral parameters over a phoneme. This new representation for the spectral envelope is then better adapted for VC, in which case, the synthesis quality is expected to be higher compared to an approach treating frames independently.

4. CONCLUSIONS & FUTURE WORK

In this work, we have proposed a new method for estimating the speech spectral envelope that simultaneously models spectral peaks and their evolution in time. We have highlighted important properties of our approach, including smooth parameter evolution and regularity within a phoneme. Furthermore, we have shown comparable performance of our analysis with that of a more traditional, cepstrum-based approach.

Given the properties of our analysis, notably the relative independence of different regions of the spectrum and the regular evolution of the spectral parameters in time, this approach to spectral envelope modeling is particularly well-suited for interpolation of spectral parameters. For example, this analysis can be used for smoothing between speech units in concatenative synthesis, by imposing further continuity constraints across phoneme boundaries. Additionally, interpolation can be used to reduce the number of parameters needed to capture the speech envelope or, in a conversion context, to generate spectral parameters in-between selected regions of phonemes that have been transformed.

Most importantly, the advantages of using this type of approach to spectral envelope analysis that models a regular evolution of spectral parameters will be most evident in a VC context, as we will examine in future work.

REFERENCES

[1] Turk, O., and Arslan, L., "Robust processing techniques for voice conversion," Computer Speech and Language 20, 2006, 441-467.

[2] Laroche J., and Dolson, M., “New phase-vocoder techniques for pitch-shifting, harmonizing and other exotic effects,” in IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, Oct. 1999.

[3] *Springer Handbook of Speech Processing*, Editors Benesty, J., Sondhi M. & Huang Y., Springer, 2008.

[4] Nguyen, B. and Akagi, M., "Spectral Modification for Voice gender Conversion Using Temporal Decomposition," Journal of Signal Processing, Vol. 11, No. 4, pp. 333-336, July 2007.

[5] Zolfaghari, P., Watanabe, S., Nakamura, A. and Katagiri, S. "Bayesian Modelling of the Speech Spectrum Using Mixture of Gaussians," in Proceedings of ICASSP '04, pp. 553-556.

[6] Nguyen, B., "Studies on Spectral Modification in Voice Transformation," Ph.D. diss, Japan Advanced Institute of Science and Technology, March 2009.

[7] Stylianou, Y. "Harmonic Plus Noise Models for Speech, combined with Statistical Methods, for Speech and Speaker Modification," Ph.D. diss., ENST, Paris, France, Jan. 1996.